

New Data Dissemination Approaches in Old Europe

Synthetic Datasets for a German Establishment Survey

Jörg Drechsler

*(Institute for Employment Research,
Germany)*

NSF-Census-IRS Workshop on Synthetic Data and
Confidentiality Protection 2009, Washington, D.C.

31. July 2009



Overview

- Background
- Summary of Results from Preliminary Studies
- The Imputation Task
- The Synthesis Task
- Disclosure Risk Evaluations
- Data Utility Evaluations
- Conclusions/Future Work



Project Background

- Yearly conducted establishment survey (IAB Establishment Panel)
- Strong demand for access from external researchers
- Only on-site and remote access possible so far
- High costs in terms of time and money
- Project-goal: Generate synthetic datasets of the survey for public release
- Project start: summer 2006

The IAB Establishment Panel

- Annually conducted establishment survey
- Since 1993 in Western Germany, since 1996 in Eastern Germany
- *Population:* All establishments with at least one employee covered by social security
- *Source:* Official Employment Statistics
- Sample of more than 16.000 establishments in the last wave
- *Contents:* employment structure, changes in employment, business policies, investment, training, remuneration, working hours, collective wage agreements, works councils

Summary of Results from Preliminary Studies (Drechsler et al., 2008)



- Find a published regression that uses only one wave of the panel
- Ask author for permission to replicate regression with synthetic data
- Generate fully synthetic datasets for a subset of the wave 1997
- Generate partially synthetic datasets for the same subset
- Evaluate data utility by comparing the regression results
- Evaluate disclosure risk

Summary of Results from Preliminary Studies (Drechsler et al., 2008)



- Generating synthetic datasets can be a useful method for SDC
- Advantages for partially synthetic datasets:
 - Higher data validity
 - Imputation models easier to set up
 - Lower risk of biased imputations
- Disadvantages for partially synthetic datasets:
 - Higher risk of disclosure
 - True values remain in the dataset
 - Only survey respondents are included
 - Careful disclosure risk evaluation necessary
- The IAB will release partially synthetic datasets for the wave 2007

The Imputation Task



- More than 250 of the close to 300 variables contain missing values
- Missing rates modest for most variables ($<1\%$ for 65.8% of the variables, only 12 variables with missing rates above 5%)
- Skewed distributions, logical constraints and skip patterns make the modeling task a nightmare
- Imputation by sequential regression
- Three imputation models:
 - linear models for continuous variables
 - logit models for binary variables
 - multinomial models for categorical variables

The Imputation Task



- Imputation models condition on all variables without skip patterns
- Some variables are dropped for multicollinearity reasons
- Multinomial imputation models are limited to 30 variables found by stepwise regression to speed up the imputation procedure
- We generate $m=5$ imputed datasets
- Runtime 3-4 weeks

The Synthesis Task



- Almost all continuous variables are synthesized
- Combination of variables that could be used for re-identification purposes (e.g. region, industry, establishment size) and sensitive variables (e.g. turnover, subsidies)
- All records are synthesized for each variable
- Several independent imputation models for each variable
- Categorical variables are synthesized using CART models (Reiter, 2005)
- $r=5$ synthetic datasets for every imputed dataset ($m \cdot r=25$ datasets)

Measures for Disclosure Risk (Drechsler & Reiter, 2008)



- disclosure risk measures based on Reiter & Mitra (JPC, 2009) and Drechsler and Reiter (PSD, 2008)

- Assumptions: - Intruder has exact information for some target records t from external databases

- target records may or may not correspond to a unit in the released data

- Calculate matching probability for each record in every synthetic dataset for every target record t

- Average over the synthetic datasets to get average matching probability

- Final risk measures are summaries of these matching probabilities

- **True match rate:** (number of correct single matches)/ N

- **False match rate:** (nb of false single matches)/nb of single matches

Assumptions for the simulation

■ Assumptions about the intruder:

- Intruder has exact information on the number of employees, the industry code, and region for a sample of establishments
- sample is a new sample from the GSSD using the same sampling design as for the panel (stratification by establishment size, region and industry code)
- No information about the generation of the data is released

■ Observation from the survey is considered a match if

- $industry_{obs} = industry_{target}$
- $region_{obs} = region_{target}$

$$nb.emp_{target} - \sqrt{sd_s(nb.emp_{target})} \leq nb.emp_{obs} \leq nb.emp_{target} + \sqrt{sd_s(nb.emp_{target})}$$

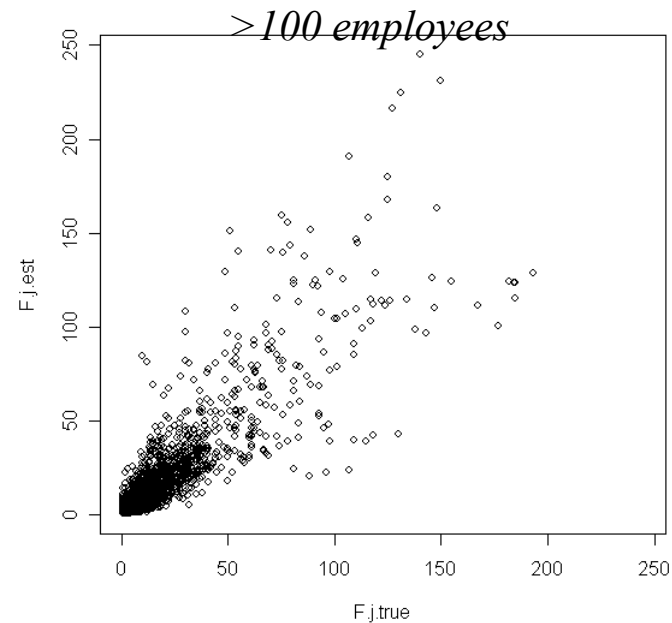
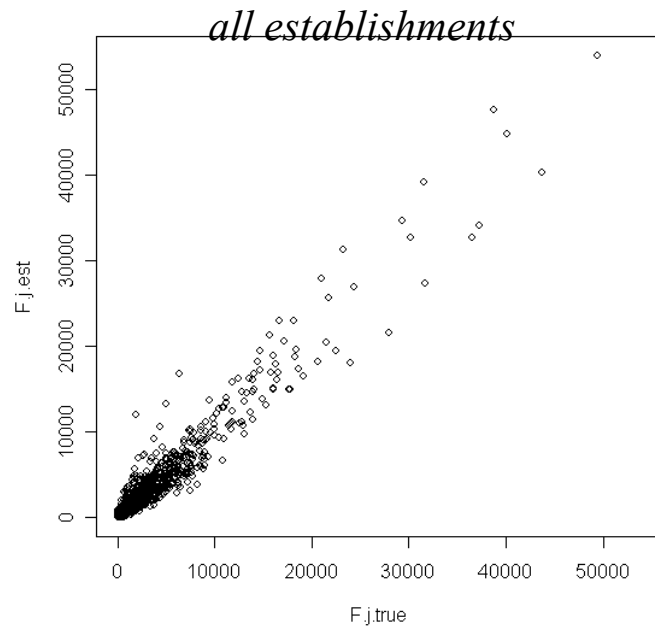
sqrt of the std. dev. for the nb of emp. in cell s

$$\sqrt{sd_s(nb.emp_{obs})}$$

■ 10 different stratification cells s for establishment size

Assumptions for the simulation

- Fit log-linear model to original data to estimate F_j
- Test goodness of fit for log-linear model by comparing with true F_j



Simulation results

■ 907 records are included in both samples

■ Probability to be included samples depends on establishment size

nb of employees covered by social security	probability to be included in both samples in %
1-4	0.79
5-9	1.78
10-19	2.53
20-49	3.85
50-99	6.73
100-199	11.94
200-499	16.52
500-999	20.89
1000-4999	31.13
>=5000	46.43

■ True match rate never exceeds 7% in any establishment size class

	F.j.true	F.j.est
True match rate in %	0.96	0.96
False match rate in %	98.76	98.76

But...

- Very large establishments can be identified by matching only on establishment size
- Intruder can ignore that region and industry code is different
- Two possible intruder scenarios:
 - Intruder ranks the synthetic data records by establishment size
 - Intruder uses nearest neighbor matching between target records and synthetic records
- Sampling probabilities close to 1 for large establishments
- We use the original data for the matching

Disclosure Risk for Large Establishments

- Results for the largest 15 establishments

<i>Original rank</i>	<i>mode of syn. rank</i>	<i>sd (rank)</i>	<i>average match rate</i>
1	1	0.20	0.96
2	2	0.20	0.72
3	3	0.00	1.00
4	4	0.00	1.00
5	5	0.00	1.00
6	6	0.00	0.88
7	7	0.00	0.64
8	8	0.82	0.56
9	9	0.68	0.44
10	10	0.77	0.32
11	11	0.40	0.84
12	12	0.45	0.56
13	13	0.58	0.56
14	14	0.65	0.68
15	15	0.61	0.76

Definition for sufficient protection

- Units are considered sufficiently protected if:
 - $sd_i(rank_{est.size}(j)) > 2$ with $i = 1, \dots, n$ and $j = 1, \dots, mr$
 - Less than 3 correct matches or $mode(\text{declared match}) \neq \text{correct match}$
- Records that fail the above criteria are replaced by new draws from a variance inflated imputation model
- We inflate the variance of the parameters in the regression model

$$\beta \sim N(\hat{\beta}, \alpha * \sigma^2 (X'X)^{-1})$$

Iterative replacement procedure

- All records that fail the protection criteria are replaced by new draws from the variance inflated imputation model with a given level of α
- Records that still fail the criteria after 10 independent draws from the model, are replaced with draws from a model with the next higher level of α
- Selected levels of $\alpha = (10, 100, 1000)$
- Overall records 79 are replaced
- Less than 10 records are replaced with $\alpha \geq 100$

Disclosure Risk for Large Establishments

■ Ranking Scenario

$\text{mode}(\text{rank}_{\text{est.size}}(i_j)) = \text{rank}_{\text{est.size}}(i_{\text{org}})$ for only 12 of the 100 largest establishments

■ The intruder never knows if declared match is correct

■ Nearest neighbor matching scenario

■ $\text{Mode}(\text{declared match}) \neq \text{correct match}$

■ Record is never identified correctly in more than 5 of the 25 synthetic datasets

	F.j.true	F.j.est
True match rate in %	0.73	0.75
False match rate in %	98.97	98.95

A quick look at the data utility

- Two Regressions suggested by colleagues at the IAB
- First regression:
 - dependent variable: part-time employees yes/no
 - probit regression on 19 explanatory variables + industry dummies
- Second regression:
 - Dependent variable: expected employment trend (decrease, no change, increase)
 - ordered probit on 38 variables + industry dummies
- Both regressions are computed separately for West and East Germany

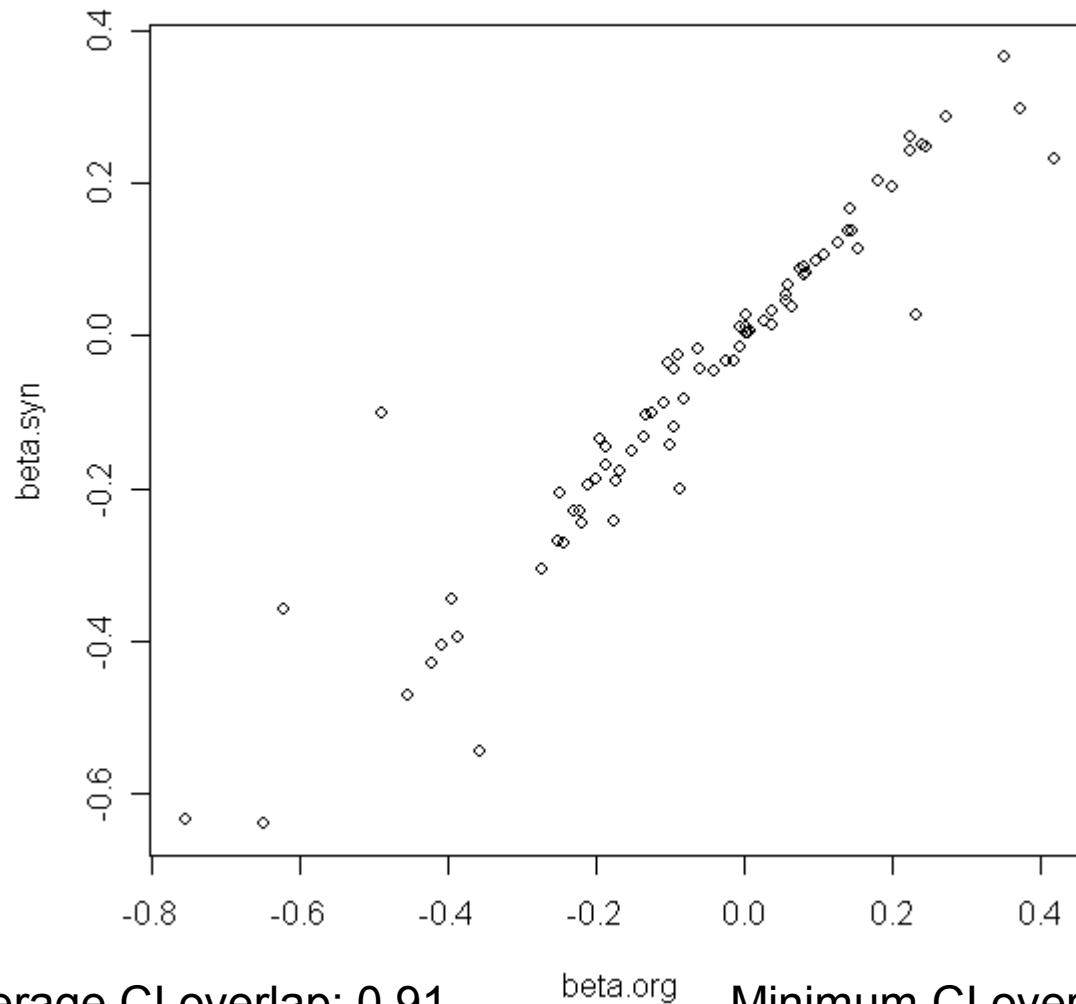
Regression results for West Germany

	<i>beta org.</i>	<i>beta syn.</i>	<i>J.k.beta</i>	<i>z-score org.</i>	<i>z-score syn.</i>	<i>CI length ratio</i>
Intercept	-0.809	-0.752	0.87	-7.23	-6.85	0.99
5-10 employees	0.443	0.437	0.97	8.52	7.99	1.06
10-20 employees	0.658	0.636	0.90	11.03	10.88	0.98
20-50 employees	0.797	0.785	0.95	13.02	12.36	1.04
100-200 employees	0.892	0.908	0.96	9.23	9.48	0.99
200-500 employees	1.131	1.125	0.99	9.99	9.87	1.01
>500 employees	1.668	1.641	0.97	8.22	8.33	0.97
growth in employment exp.	0.010	0.006	0.98	0.18	0.12	0.99
decrease in emp. expected	0.087	0.100	0.96	1.11	1.27	1.00
share of female workers	1.449	1.366	0.73	17.63	18.71	0.89
share of employees with university degree	0.319	0.368	0.91	2.18	2.59	0.97
share of low qualified workers	1.123	1.148	0.93	12.17	11.87	1.05
share of temporary employees	-0.327	-0.138	0.75	-1.74	-0.71	1.05
share of agency workers	-0.746	-0.856	0.88	-3.09	-4.24	0.84
employment in the last 6 month	0.394	0.369	0.87	8.33	7.82	1.00
dismissal in the last 6 months	0.294	0.279	0.92	6.38	6.03	1.00
foreign ownership	-0.113	-0.117	0.99	-1.33	-1.38	0.99
good or very good profitability	0.029	0.033	0.98	0.72	0.82	0.99
salary above collective wage agreement	0.020	0.031	0.95	0.35	0.54	0.99
collective wage agreement	0.016	0.007	0.95	0.31	0.13	0.97

Regression results for East Germany

	<i>beta org.</i>	<i>beta syn.</i>	<i>J.k.beta</i>	<i>z-score org.</i>	<i>z-score syn.</i>	<i>CI length ratio</i>
Intercept	-0.712	-0.742	0.93	-6.42	-7.21	0.93
5-10 employees	0.266	0.257	0.96	4.81	4.53	1.03
10-20 employees	0.416	0.399	0.93	6.94	6.76	0.99
20-50 employees	0.542	0.532	0.96	9.18	8.72	1.04
100-200 employees	0.757	0.808	0.86	8.02	8.47	1.01
200-500 employees	0.971	1.013	0.91	8.25	8.57	1.00
>500 employees	1.401	1.422	0.98	5.69	5.66	1.02
growth in employment exp.	-0.041	-0.040	1.00	-0.73	-0.73	1.00
decrease in emp. expected	0.035	0.040	0.98	0.44	0.50	1.00
share of female workers	1.006	1.041	0.88	12.63	14.93	0.88
share of employees with university degree	0.221	0.197	0.95	1.86	1.76	0.95
share of low qualified workers	0.976	1.042	0.87	8.44	7.84	1.19
share of temporary employees	-0.049	0.049	0.84	-0.31	0.34	0.91
share of agency workers	-0.176	-0.232	0.94	-0.73	-1.08	0.89
employment in the last 6 month	0.230	0.210	0.89	4.95	4.55	1.00
dismissal in the last 6 months	0.301	0.295	0.97	6.43	6.35	0.99
foreign ownership	-0.176	-0.176	1.00	-1.83	-1.84	1.00
good or very good profitability	0.097	0.097	1.00	2.35	2.37	1.00
salary above collective wage agreement	0.080	0.086	0.98	1.04	1.10	1.01
collective wage agreement	0.097	0.069	0.86	1.87	1.36	0.98

results for the second regression



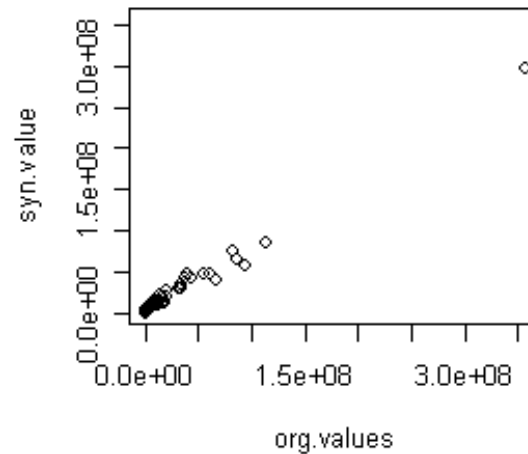
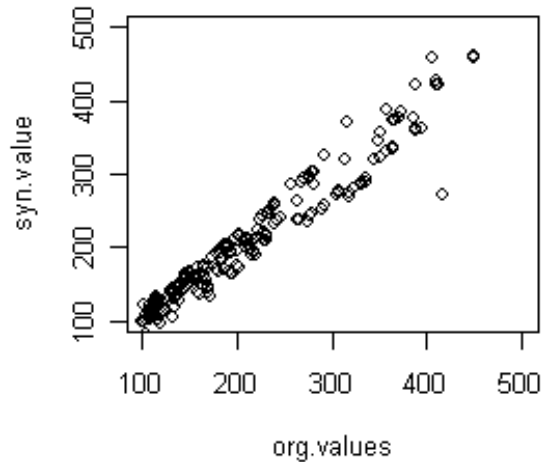
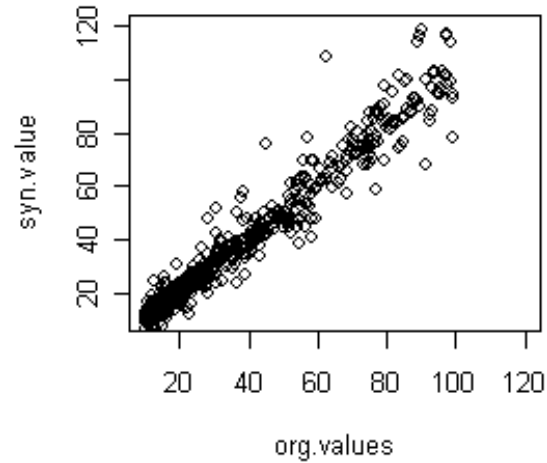
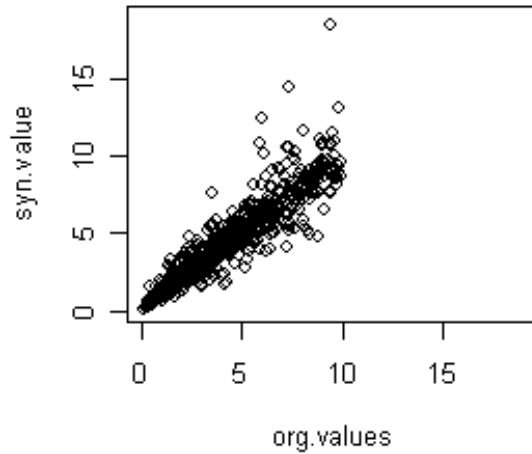
■ Average CI overlap: 0.91

Minimum CI overlap: 0.61

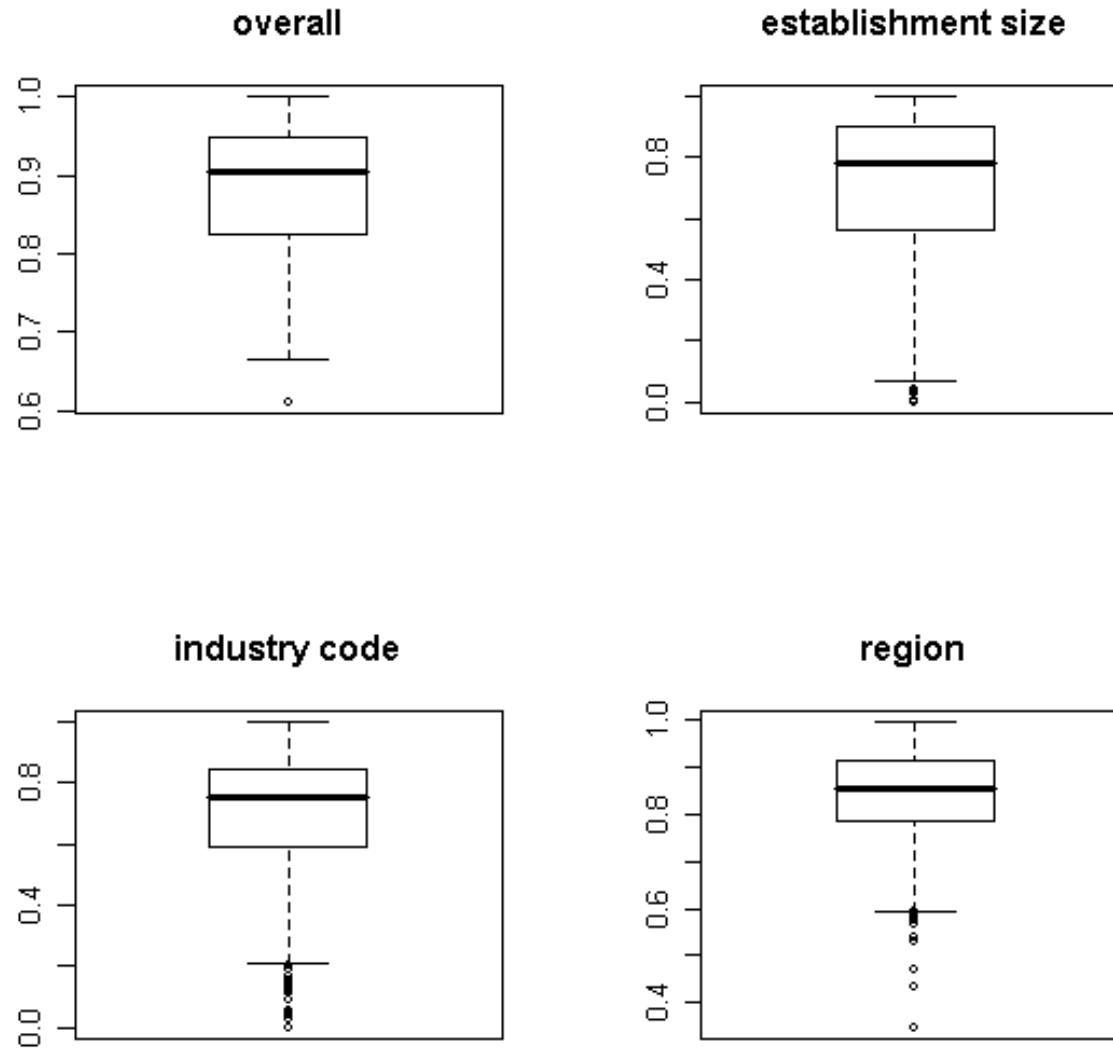
Descriptive Comparison

- Compare the unweighted means for all continuous variables for different subgroups
- Subgroups:
 - establishment size (10 categories defined by quantiles)
 - industry dummies (17 categories)
 - region (16 categories)
 - overall mean
- All categories with at least 200 observations above zero are compared

True mean vs. synthetic mean



CI overlap for different subgroups



Conclusions

- Synthetic datasets provide a high level of disclosure protection
- Synthetic datasets offer a high level of data utility
- Datasets almost ready for release
- Always ways to improve data quality
- Interaction with the user

Future Work

- Provide metadata for the user
- Long term goal: release complete longitudinal data

Thank you for your attention

The Synthesis Task



- Data quality in publicly available databases lower than expected
- Re-identification experiment based on probabilistic record linkage
- Only a small fraction of records in the original data could be correctly identified
- Results helped in the discussions with the DRB
- Decisions on which variables to synthesize in close coordination with the department responsible for the establishment panel and *infratest* (responsible for conducting the survey)
- Almost all continuous variables are synthesized

Measures for Disclosure Risk (Drechsler & Reiter, 2008)



- disclosure risk measures based on Reiter & Mitra (JPC, 2009)
- Compute probabilities of reidentification for each record j ($j=1, \dots, n$) in the released dataset
- Assumptions: - Intruder has exact information for some target records t from external databases
 - target records may or may not correspond to a unit in the released data
- Let t_0 be the unique identifier for the target record
- Let d_{j_0} be the identifier for record j in the released data D , $j=1, \dots, s$
- Intruders goal: match if $t_0 = d_{j_0}$; don't match if $t_0 \neq d_{j_0}$

Measures for Disclosure Risk II

- Let J be a random variable with

$$J = \begin{cases} j & \text{for } d_{j0} = t_0 \text{ and } j \in D \\ s+1 & \text{for } d_{j0} = t_0 \text{ and } j \notin D \end{cases}$$

$$\Pr(J = j \mid t, D, M)$$

with: D set of released synthetic datasets
 M any additional information about the generation of D

- Intruder does not know actual values in Y_{rep}
- Integrate over its possible values
- Monte Carlo approach to estimate $\Pr(J = j \mid t, D, M)$

Example

- let age, race, and sex be the only quasi-identifiers in a survey
- Agency releases no information about the imputation models
- Intruder seeks to identify a white male aged 45 and **knows the target is in the sample**
- Intruder would match on age, race and sex
- Average matching probability

$$p_{match,i} = \Pr(J = j \mid t, D, M) = (1/m) \sum_k (1/N_k) I_i$$

with N_k = nb of records that fulfill the matching criteria in dataset k

I_i = 1 if record i is among the N_k records, 0 otherwise

m = number of synthetic datasets

- Probability that target record is not in the sample: zero

Example II

- Intruder seeks to identify a white male aged 45 and **does not know the target is in the sample**
- Replace N_k with F_t , the number of records in the *population* that match on age, race and sex
- Estimate F_t with a log-linear model (Elamir & Skinner, 2006)
- Fit log-linear with original data (conservative) or released data
- If $N_k > F_t$, intruder picks one of the matching records at random

$$p_{match,i} = \Pr(J = j | t, D, M) = (1/m) \sum_k \min(1/F_t, 1/N_k) I_i$$

with F_t = nb of records that fulfill matching criteria in the population N_k
 = nb of records that fulfill matching criteria in dataset k
 I_i = 1 if record i is among the N_k records, 0 otherwise
 m = number of synthetic datasets

Measures for Disclosure Risk III

- Summaries of the average match risks
- Intruder selects record j with highest value of $Pr(J=j|t,D,M)$
- Further definitions:
 - c_j = number of records with $\max(p_{match,i})$ for target t_j
 - I_j = 1 if true match is among the c_j units, 0 otherwise
 - K_j = 1 if $c_j I_j=1$, 0 otherwise

3 disclosure risk measures

- Expected match risk $\sum_j (1/c_j) I_j$
- True match risk $\sum_j K_j$
- True match rate $\sum_j K_j / \sum_j (c_j = 1)$

Analytical validity

- Compute the (unweighted) average number of employees by industry using the original data and the synthetic data
- Calculate the confidence interval overlap as suggested by Karr et al. (2006)
- Measure the overlap of CIs from the original data and CIs from the synthetic data
- The higher the overlap, the higher the data utility

$$J_k = \frac{1}{2} \left[\frac{U_{over,k} - L_{over,k}}{U_{orig,k} - L_{orig,k}} + \frac{U_{over,k} - L_{over,k}}{U_{syn,k} - L_{syn,k}} \right]$$

